Disclosing the Genetic Footprint of the Bacterial Ecotype Inhabiting the Gut of Homeothermic Hosts through Comparative Metagenomics Studies

Joana Lima
Biologia
2013

Orientador
Professor Doutor Fernando Tavares
Faculdade de Ciências da Universidade do Porto

Disclosing the Genetic Footprint of the Bacterial Ecotype Inhabiting the Gut of Homeothermic Hosts through Comparative Metagenomics Studies

Dissertação submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Biodiversidade, Genética e Evolução.

Local de realização:

Departamento de Biologia – Faculdade de Ciências da Universidade do Porto

Centro de Investigação em Biodiversidade e Recursos Genéticos – CiBio

Orientador:

Professor Doutor Fernando Tavares

Centro de Investigação em Biodiversidade e Recursos Genéticos – CiBio

Faculdade de Ciências da Universidade do Porto

"Nothing in Biology Makes Sense Except in the Light of Evolution."

- Theodosius Dobzhansky

Abstract

Bacteria are excellent models to study evolutionary processes and ecological adaptability – bacterial genomes are small-sized and therefore nowadays simple to sequence. *Escherichia coli* species in particular is one of the most ecologically versatile *taxon* known today – it is an important member of the human microbiota, while it can also be found in outer environments such as environmental waters or the soil, and it has also been widely exploited, becoming a laboratory workhorse. The introduction of next generation sequencing approaches has caused a rapid increase in the number of completely sequenced genomes available. More each day, the amount of genetic data and bioinformatics tools allows a deeper and more thorough understanding of the flexibility and functionality of genomes, while improving methods for data storage, organization and analysis.

In the past, research on the impact of bacteria on the human gastrointestinal tract was mainly focused on pathogenic organisms and on the way they caused diseases. Nowadays, importance is being given to the human microbiota, microflora or normal flora – the vast set of bacterial organisms that live in peaceful coexistence with their hosts. In this work, the main goal is to identify the genetic features shared by non-related organisms belonging to ecologically similar microbiotas – and ultimately to disclose the genetic footprint of the bacterial ecotype capable of surviving in the gut of homeothermic hosts.

Using EDGAR 1.2, a comparative analysis between five *Escherichia coli* strains was performed and a set of genes was targeted as putatively related to the adaptation and survival of bacteria within the gastrointestinal tract of homeothermic animals. Blasting each of those genes against the nucleotide collection of NCBI allowed the introduction of multiple non-related organisms to this study, reported also as constitutive parts of the group of bacteria capable of surviving within the gut of homeothermic hosts as *Shigella* spp., *Klebsiella* spp., or *Salmonella* spp. Two thresholds were applied on the Blast result-set in order to filter out organisms and CDSs not related with the survival of bacteria within the gut of homeothermic hosts, resulting on a list of twenty genes putatively responsible for metabolic functions related to adaptation to this habitat.

Key-words: *Escherichia coli*, comparative genomics, bioinformatics, microbiota, microbiome

Resumo

As bactérias são excelentes modelos para estudar processos evolutivos e a adaptabilidade ecológica – os seus genomas são pequenos e por isso a sua sequenciação é simples. A espécie *Escherichia coli* em particular é uma das divisões mais conhecidas pela sua versatilidade em termos de nichos colonizados – é um membro importante do microbiota humano, podendo também ser encontrada em ambientes exteriores como águas ambientais ou solo, e tem sido também extensivamente explorada no laboratório, tornando-se um destacável objecto de estudo. O aparecimento de técnicas como a *next generation sequencing* ou o *whole genome shotgun sequencing*, por exemplo, causou um aumento muito rápido no número de genomas sequenciados disponível hoje em dia. Cada vez mais, a quantidade de dados genómicos e ferramentas bioinformáticas disponível permite um mais profundo e pormenorizado conhecimento da flexibilidade e funcionalidade dos genomas, pelo melhoramento de processos de armazenamento, organização e análise da informação genética.

No passado, os estudos sobre o impacto das bactérias no tracto gastrointestinal humano eram principalmente orientados para a exploração de organismos patogénicos e da forma como estes causam doenças. Hoje em dia, mais importância tem vindo a ser dada ao microbiota, micro flora ou flora normal do ser humano - o vasto conjunto de bactérias que coexiste pacificamente com o seu hospedeiro. Neste trabalho, o principal objectivo é identificar quais são os genes partilhados por organismos não relacionados pertencentes a nichos ecologicamente similares – genes partilhados por bactérias que tenham sido maioritariamente isoladas do tracto gastrointestinal de animais homeotérmicos.

Usando o EDGAR 1.2, uma análise comparativa entre 5 genomas de estirpes de *Escherichia coli* foi efectuada e um conjunto de genes surgiu como putativamente relacionado com a adaptação e sobrevivência das bactérias do tracto intestinal de animais homeotérmicos. O blast de cada um dos genes pertencentes a este conjunto permitiu a introdução no estudo de organismos não relacionados com *E. coli*, sendo

estes também parte do grupo de organismos capazes de sobreviver dentro do tracto gastrointestinal de hospedeiros homeotérmicos, como por exemplo *Shigella* spp., *Klebsiella* spp. or *Salmonella* spp. Foram aplicados dois filtros sobre o resultado do Blast para impedir a entrada de organismos não relacionados com este tipo de nicho.

Este trabalho propõe desvendar a pegada genética de ecótipos bacterianos adaptados à sobrevivência e adaptação dentro do tracto gastrointestinal de hospedeiros homeotérmicos usando ferramentas bioinformáticas fiáveis e rápidas.

Palavras-chave: *Escherichia coli*, genética comparativa, bioinformatica, microbiota, microbioma

List of Tables

Table 1 ……..……..……..30

Table 2 ……..……..……..37

Table 3 ……..……..……..42

Table 4 ……..……..……43-45

Table 5 ……..……..……..48

List of Illustrations

Figure 1 30

Figure 2 35

Figure 3 38

List of abbreviations

HGT: Horizontal Gene Transfer

DNA: Deoxirribonucleic Acid

RNA: Ribonucleic Acid

AARD: Average Annual Rate of Decline

WHO: World Health Organization

MDG: Millenium Development Goals

EDGAR: Efficient Database framework for comparative Genome Analyses using BLAST score Ratios

NCBI: National Center for Biotechnology Information

BBH: Bidirectional BLAST Hits

BLAST: Basic Alignment Search Tool

SRV: Score Ratio Value

LSW: Lowest Scoring Window

E-value: Expect value

S: Score

C. g.: Complete Genome

CMR: Comprehensive Microbial Database

MBGD: Microbial Genome Database

COG: Cluster of Orthologous Groups of Proteins

CAMP: Cationic Antimicrobial Proteins

VGT: Vertical Gene Transfer

Table of Contents

Abstract....6

Resumo....8

List of Tables....10

List of Illustrations....11

List of abbreviations....12

Introduction....15

1. The Gut Microbiota, a Paradigm of Bacterial-Animal Interactions....15
2. The Central Dogma of Molecular Biology....17
3. Genetic variability, a driving force of adaptation....17
4. Bioinformatics....19
5. Genetic footprint in extreme environments....21
6. Genetic footprint on complex environments....23
7. *Escherichia*....25
8. Environmental variables....26
   - 8.1. *Nutrient availability*....26
   - 8.2. *Temperature*....27
   - 8.3. *pH*....27
   - 8.4. *Water activity (a_w)*....28
9. Objectives....28

Materials and Methods....30

1. Selection of the Working Model....30
2. Data Mining....30
3. Application of the model....30

Results and Discussion....33

EDGAR analysis disclosed the genes shared by commensal *E. coli* strains Se11 and SE15.....33

The Baas-Becking Hypothesis.....34

The Venn Diagram 36

References.....51

Introduction

Bacteria are excellent models to study evolution and ecological adaptability. Besides the capability to colonize a multitude of environments - each characterized by its own unique set of physical, chemical and ecological properties - bacteria also exhibit a short generation time and a highly plastic genome. Bacterial organisms are distributed almost ubiquously throughout the multiple habitats Earth provides, yet, microbial biogeographic questions do not appear to have had great impact so far. The limitless vocation of bacteria to colonize all these different niches reveals that bacteria have evolved in a way to survive and adapt to the widest range of conditions. The successful bacterial adaptability is mainly understood by the plasticity of their microbial genomes and their uneven distribution across habitats. With an average of 140 to 13,000 genes, microbial genomes are characterized by being subjected to rearrangement processes, such as deletions, insertions, duplications and shuffling processes, frequently due to dedicated entities and mechanisms such as mobile genetic elements or HGT[1], resulting in gene loss and acquisition or even in the translocation of whole genomic islands, which can either increase or decrease their ecological adaptability as a species (Altermann 2012).

1. The Gut Microbiota, a Paradigm of Bacterial-Animal Interactions

Bacterial evolution and adaptation to different environments has been the focus of attention of numerous researchers. The study of bacteria colonizing environments characterized by extreme biological, physical or chemical conditions, for example, extreme salinity levels or presence of closely-related competitors, provides very important reference points which aid in the understanding of the bacterial mechanisms necessary to cope with environmental challenges. In this regard, one major goal over the years has been to disclose the different coevolutionary processes of bacterial communities within animal hosts. From the numerous examples of hot spots for

[1] Horizontal Gene Transfer

bacterial-animal interactions, the gut microbiota[2] of homoeothermic animals (the human gut microbiota in particular) has been extensively studied as a paradigmatic model to address the importance that bacteria have had in shaping up a partnership acknowledged as extremely complex. It has been estimated that at least 500 – 1,000 different microbial species exist in the human gastrointestinal microbiota (Cabral 2010). The idea that numerous bacteria colonizing the gut have an exclusively pathogenic behavior, i.e. bacteria benefiting from the host while negatively affecting it, has changed dramatically. In the last decade, numerous studies have been published emphasizing the positive role gut microbiota has in the host's metabolism, physiology, resistance to diseases and even behavior by metabolizing human indigestible biomolecules (Blaut and Clavel 2007; Sekirov et al. 2010), modulating the host immunological system (Gibson and Roberfroid 1995; Guarner and Malagelada 2003; Blaut and Clavel 2007; Kurokawa et al. 2007; Sekirov et al. 2010); producing and releasing antimicrobial compounds related with inhibition of growth of pathogenic bacteria and, at the same time, stimulating the host's immune system to produce antimicrobial compounds (Sekirov et al. 2010); influencing reproductive behavior in both vertebrate and invertebrate animals (Ezenwa et al. 2012) and even modulating human brain physiology (Collins et al. 2012). The intestinal microbiota consists on a vast microbial community that lives generally in a symbiotic relationship with the host. Yet, the gut microbiota might also negatively impact the host. Studies have shown that gut microbiota may be related to several diseases, namely the inflammatory bowel disease (IBD) (Sartor 2008; Honda and Takeda 2009; Sekirov et al. 2010), HIV (Hofer and Speck 2009; Sekirov et al. 2010) and cancer (Martin et al. 2004). Altogether these studies contributed to shed some light on the panoply of roles played by microorganisms in the ecosystems, and also to call attention to the enrichment of determined bacterial *taxa* across different habitats, which suggest a high-level specialization by many different strains. For example, 90% of the bacterial species in termite guts are not found elsewhere (Hongoh 2010; McFall-Ngai et al. 2013).

[2] Any group of microbial taxa inhabiting the same ecosystem, Oshima.et al. (2008).

2. The Central Dogma of Molecular Biology

The fitness of each prokaryotic cell to a certain environment is always a consequence of its genotype. Each step in any metabolic pathway is controlled by one or more proteins. The central dogma of molecular biology deals with the detailed residue-by-residue transfer of **sequential information**. It states that such information cannot be transferred from protein to either protein or nucleic acid (Crick 1970).This idea could be stated as: the information retained in the coding DNA is transcript in a specific way into an RNA molecule, and from there, it will be specifically translated into a protein. Each codon will, upon reading, originate one amino acid. Although various codons hold information for the production of the same amino acid (degeneracy of the genetic code), each codon is specifically related to the production of a specific amino acid. The direction is DNA : RNA : protein and never the other way around. A gene is a set of nucleotides that once transcript and translated will produce a protein. Proteins constitute the primary definition of phenotype. These molecules are unique and function-specific, so each one plays each part in a metabolic pathway. Each protein acts on the product of a reaction catalyzed by the previous one. Accordingly, each microbial community is expected to present a determined set of genes and to be identifiable by the exclusive presence or enrichment of determined sets of sequences.

3. Genetic variability, a driving force of adaptation

In 1953, DNA's structure was disclosed by Watson and Crick's work, which enabled some alterations in the way biology is done. Nowadays, DNA is the central entity in the biological sciences world as it is the hereditary material in humans and almost all organisms. To study this field allows scientists to understand the blueprint for building a person, a bacterium or other organisms, to understand relations between organism and environment and between the organisms themselves. This knowledge is already having a major impact in the fields of medicine, biotechnology, and the life sciences.

The functional flexibility of bacterial genomes related to the differential gene expression (Smoot et al. 2001; Revel et al. 2002) and variation in gene content (Brosch et al. 2001) is achieved by horizontal gene transfer (Ochman et al. 2000), gene duplication

events (Jordan et al. 2001) or even by other types of mutation processes, which also contribute for the genetic variation of bacterial genomes, as it is the case of point mutations, which can lead to silent mutations, causing no change in the final protein. Silent mutations don't have implications in terms of survival or fitness. Yet, when there are variations to the produced proteins, there could be implications to the metabolic pathways of the cell: the expected protein could be produced in a different amount, provide a different function, or provide the same function but not as well as the original protein. This means that there could be differences in the fitness of the cell to its habitat. In the same niche it is expected that multiple variations to the original organism appear spontaneously throughout time. Variations will rise and fall, perishing or surviving, according to their fitness to the environment. Regarding the DNA itself as a self-perpetuating entity, organisms will appear as empty shells that will or will not pass their DNA onto the next generation. According to this perspective, fitness becomes the sole means to success. One is fit when one is able to survive and reproduce, passing on one's genes to the following generation. Yet, fitness is not a static place or state. It could be described as the degree of success in response to the distribution of resources and competitors at a determined time or period of time.

Although new genes can be originated through duplication of existing sequences, followed by diversification, the most common way to acquire new functions is by the transfer of genetic material from unrelated organisms (Medini et al. 2005). Organisms inhabiting the same environment contribute to a so-called gene pool, loosing and gaining genes from it by three main HGT processes: (i) transformation, when genetic material can be taken up directly from the environment; (ii) transduction, when the DNA is delivered by a virus and (iii) conjugation, when DNA is directly exchanged between cells (Medini et al. 2005). All of these processes imply that the cell that provided the DNA uptaken by other cell has been in that same environment where the uptake of the genetic material took place - contact with the same gene pool – they have inhabited in the same environment, even if not at the same time. McFall-Ngai, M. *et al.* (2012), states that, not surprisingly, many animal genes are homologs of bacterial genes, mostly derived by descent, but occasionally by gene transfer from bacteria (Keeling and Palmer 2008; McFall-Ngai et al. 2013).

Bacterial organisms hold in their genomes a determined set of genes responsible for the basic maintenance of the prokaryotic cell. At the same time, bacterial lineages

maintain a genetic stability within a determined set of shared genes based on VGT from generation to generation, which may be blurred by HGT events. Independent lineages of organisms belonging to the same community, and sharing an environment and all the challenges it presents to their survival, are thought to share the genetic material responsible for the adaptation to the specific set of environmental conditions the habitat presents. Particularly, bacterial communities inhabiting the gastrointestinal tract of homeothermic animals ultimately hold in their individual genomes a determined set of genes responsible for their adaptation and survival in this habitat, while at the same time maintain a certain degree of flexibility which allows them to cope with short-time variations such as dietary changes, ingestion of antibiotics, variable number and type of competitors and other challenges the gut of homeothermic animals represents.

4. Bioinformatics

The conjugation of biology with computer science has given rise to new fields like bioinformatics and computational biology. Since bioinformatics exists, methods for storing, retrieving, organizing and analyzing biological data have greatly improved. Computing contributed not only to the raw capacity for processing and data storage, but also to the mathematically-sophisticated methods to achieve the results. Techniques such as image processing or network calculations with multiple algorithms allow for different, richer and more precise extraction of information from large amounts of raw data. Nowadays, there is almost too much raw data available: sequences of genes, whole genomic sequences, uncharacterized DNA sequences (i.e. sequences for which biological meaning stays unknown) are present in very large databases, which makes it impossible to analyze manually, because the scenario becomes very cloudy to discover all distribution patterns and relationships between organisms, genomes, populations and ecosystems. From 1953 until today, there was a dramatic change in the way data is mined in biology and in the nature of that same data. Switching from measuring values of variables whose distribution was continuous to observing nucleic acids and amino acids, whose distribution of values is discrete is an important step. It is now possible to characterize an organism, a population or even an ecosystem as a coherent genomic identity with more exactitude than ever.

Computational biology has mode possible to create databases larger up to a point that we had never been before (Medini et al. 2005; Oshima et al. 2008). But, although we have the data, we don't have the means to analyze it fully and correctly. Some believe that investing in the 'omics' technologies, such as genome-wide association studies, and the cataloging of new genomes will, all on their own, be sufficient for us to make sense of the biological complexity we can now measure (Friend and Norman 2013). Until recently, scientists were mostly focused on studying the function and organization of each gene, and the role of the protein it was responsible for, or even in its relevance to the development of the organism it belongs to. But a genocentric approach is limited to a single data dimension and is unable to provide a complete enough context to see and understand a biological system in its entirety (Friend and Norman 2013). Friend, S. et al (2013) states that like one frame in a 200.00-frames movie, a single biological reading is a frozen snapshot of a complex living system and a crippled approach to understanding the story of how biology works. The fourth dimension – time – is not accounted for.

Today, bioinformatics is an applied science. Computer softwares permit new and more precise inferences from the data archives of modern molecular biology, to make connections among them, and to derive useful and interesting predictions. Now it is possible to channel all the work through an interface to the web. A serious problem with the web is its volatility. Sites come and go, leaving trails of dead links in their way. Yet, this tool allowed for the development of a whole web of knowledge and information, which enables faster and stronger-based scientific work.

Bacterial genomic databases are continuously growing and each record is becoming more complete. There are multiple software's freely available on the web, and one can use them when investigating genomes and their evolutionary pathways or when searching for a statistical correlation between phenotype and genotype. Independently of the specific goal of the work, datasets provided by bacterial genomic databases are nowadays the workhorse of many biologists, allowing comprehensive examinations in a short period of time and avoiding the limitations typically found in laboratory works, as the inability to cultivate the majority of organisms in bacterial communities (Zogg et al. 1997), or even the economical investments involved in the isolation and sequencing of the genetic material.

5. Genetic footprint in extreme environments

Biology fields of genetics and molecular biology used to be mainly focused on studying a single organism, a single gene or even a single metabolic pathway. Nowadays, scientists are more interested and closer to integrative perspectives than ever: to relate the genome to the individual's characteristics and this to the surrounding environment provides a clearer view of the informational web formed by all organisms in their habitats, supported in their hereditary material. When different data from different sources converge, it is possible to identify faster and with more accuracy any *taxon*. The identification of any organism can be done using its genetic material – the whole genome of an individual is characteristic of that same individual and can therefore be used to identify it, yet, there is no need to use the whole molecule to do it, as the majority of the information will be shared with closely-related organisms due to VGT[3]. Regarding an organism as if it were a unique set of DNA sequences facilitates the perspective of a niche holding a set of DNA unique assemblages. A niche is characterized by its environmental conditions. It may be occupied by every organism able to survive to its environmental conditions. Accordingly, a niche is also identifiable by the unique set of DNA sequences it holds, shared by sub-sets of organisms or even by all organisms found there, and can therefore be characterized by it. This way, there is no need to use the whole genome to characterize an individual, neither to use the whole metagenome to characterize a niche but only the unique or shared features, depending on the proposed problem.

If the focused set is a niche strongly characterized by any environmental factor (i. e. extremely high temperatures or extremely low pH), then all the organisms in that environment are capable of surviving to that extreme condition, and so, studies regarding the metagenomics of this type of niches can be pre-oriented from the start: the existence of that extreme characteristic enables the scientist to more directly target a group of genes that might characterize the niche. Not all organisms in the same niche use the same survival strategies to overcome an environmental challenge. Independent lineages of organisms facing the same set of environmental factors may develop different survival strategies, as they already differ in their basic hereditary

[3] Vertical Gene Transfer

patrimony. Even if microorganisms from independent lineages present the same strategy to survive to a determined environmental challenge, they may have developed a different approach to survive another challenge. Furthermore, lineages presenting the same molecular strategies to overcome the same challenge may present different DNA sequences responsible for the phenotypical characteristics that makes them survive. These organisms, part of an unique set, share characteristics that enable them to survive the same set of challenging factors, but one cannot disintegrate: this determined and targeted set of genomic sequences is a constitutive portion of multiple and more integrative sets, which means that some of the characteristics present here will also be present in other places, even if the considered niche is completely isolated from the rest.

Searching for genes putatively associated with the survival of an organism in a determined environment is not a simple quest. HGT events give organisms in general, bacteria and archeae in particular, the possibility of introducing ready-made genes or operons in their genomes. If organisms from independent lineages share a habitat, HGT will be the reason they share genes. The genetic sequences uptaken from the shared genetic pool are usually characterized by their GC content. Because closely-related lineages exhibit similar values of %GC, genetic fragments integrated in genomes of unrelated organisms will generally present a different %GC, making the area easily identifiable. If they are from the same lineage, the difficulty in identifying the transferred genes or genomic regions will increase drastically, as this criterion will not be applicable in most cases.

The geneset that includes the sum of the genes in each organism will definitely be unique to the targeted environment, and therefore one could use this set of genes to calculate the genomic footprint of that environment. Moreover, the relative abundances of each genetic sequence within the ecosystem will provide information on which genes are persistently maintained in the population, being putatively more advantageous than genes scarcely present.

6. Genetic footprint on complex environments

When the targeted ecosystem is characterized by an extreme value of any factor – high salinity or low pH – the type of geneset that allows for the survival of organisms is predictable from the start. This kind of extreme conditions lead to the stringency of a great number of organisms and therefore do not allow for the variability of these niches to increase much. Organisms present in this kind of environments, characterized by extreme conditions, obviously survive to the extreme factor they are subjected to. Being so, in their genomes it will exist regions responsible for their survival, and these regions include information regarding the survival to that environmental specificity. But what happens when the targeted environment is a complex ecosystem and it is not strongly characterized by any extreme value? A complex ecosystem would be characterized by the absence of environmental restrictions, and that almost freely allows the entrance of new organisms, and with them new genes and increased variability. Yet, one cannot forget that even not being characterized by any extreme condition, these environments are indeed characterized by a set of environmental conditions which will or will not allow for the survival of microorganisms. This means that, when focusing on genes that could be associated to the survival of organisms in any complex environment there is no pre-targeted genomic region, because there is no characteristic recognized as definitive of the niche. Some examples of complex environments are the human gut, lake waters or even some kinds of soil. This kind of niche usually contains great variation and quantity of nutrients; temperatures are stable and not extreme (no less than 0, no more than 45ºC), and pH is close to neutrality.

The human gut specifically constitutes an interesting niche, as it presents the next level of complexity. The difficulty in this scenario resides mainly on the lack of factors that putatively would limitate the survival of microbial species. Also, there are variations in pH, temperature, nutrient availability and other critical factors throughout the human gut's length. Regardless, it is one of the best characterized niches so far. There is a high quantity of available information regarding microbial communities inhabiting here, pH variations throughout the gut, and a large quantity of microbial CDSs available on the web. Enzymes, nutrients available and regulatory cycles are also well known and characterized.

One of the enteric groups of organisms most studied until today is the group of coliforms. The work of Leclerc et al. (Leclerc et al. 2001) clarified the diversified roles that coliforms have in the environment and the real meanings of the tests on total and fecal coliforms. It was shown that *Enterobacteriaceae* encompass three groups of bacteria with very different roles in the environment. Group I includes only *Escherichia coli*. Since this species usually does not survive for long periods outside the intestinal environment, it was considered a good and reliable indicator of fecal pollution (both animal and human). Group II, the ubiquitary group, encompassed several species of *Klebsiella* (*K. pneumoniae* and *K. oxytoca*), *Enterobacter* (*Enterobacter cloacae* subsp. *cloacae*, *E. aerogenes*) and *Citrobacter* (*C. amalonaticus*, *C. koseri* and *C. freundii*). These bacteria live in the animal and human gut, but also in the outer environment, and are commonly isolated from the soil, polluted water and plants. Therefore, their presence in polluted waters does not necessarily indicate fecal ontamination. Finally, Group III was composed of *Raoultella planticola*, *R. terrigena, Enterobacter amnigenus* and *Kluyvera intermedia* (*Enterobacter intermedius*), *Serratia fonticola*, and the genera *Budvicia, Buttiauxella, Leclercia, Rahnella, Yersinia*, and most species of *Erwinia* and *Pantoea*. These bacteria live in fresh waters, plants and small animals. They grow at 4 ° C, but not at 41 ° C. They are not indicators of fecal pollution, and they can be detected in the total coliform test.

Regarding commensal organisms, each microbial community is specifically related to a certain host, evolving throughout each individual's lifetime and being susceptible to both exogenous and endogenous modifications (Sekirov et al. 2010). Microbial communities in the vertebrate gut respond to the host's diet over both daily and evolutionary time scales, harbouring animals with the flexibility to digest a wide variety of biomolecules and cope with and even flourish under conditions of diet change (Ley et al. 2008; Kau et al. 2011; Muegge et al. 2011). Furthermore, the gut microbiota adapts to changing diets and conditions not only by shifting community membership but also by changing gene content via HGT events (McFall-Ngai et al. 2013). More generally, human-associated bacteria have been shown to have a 25-fold higher rate of gene transfer than do bacteria in other environments, which highlights the important role of gene transfer in host associated bacterial communities (Smillie et al. 2011). Although animals and bacteria have different forms and lifestyles, they recognize one another and communicate in part because, as described above, their

genomic “dictionaries” share a common and deep evolutionary ancestry (McFall-Ngai et al. 2013).

Terrestrial environments often have broad, short-term (daily) and long-term (seasonal) fluctuations in temperatures. It is in these habitats that endothermy (maintaining a constant body temperature by metabolic means) evolved as a shared characteristic in birds and mammals. Most enteric bacteria of birds and mammals have optimum growth around 40 ºC, suggesting the possibility that this trait resulted from coevolution of these bacteria with their homeothermic hosts. The reciprocal may also be true, i. e., an animal’s microbial partner may have played a role in selecting for the trait of endothermy, for example by making something available for the host at a certain temperature, which could provide it with a selective advantage over hosts with a different microbiota (McFall-Ngai et al. 2013). Constant high temperature speeds up bacterial fermentation, providing rapid and sustained energy input for the host. These benefits are evident when comparing conventionally-grown to germ-free mammals, which require one-third more food to maintain the same body mass (Backhed et al. 2004). Keeping their microbes working at optimum efficiency likely offered a strongly positive selection pressure for the evolution of genes associated with the trait of endothermy in birds and mammals (McFall-Ngai et al. 2013). The interwining of animal and bacterial genomes is not just historical: by copting the vastly more diverse genetic repertoire present in its bacterial partners (Lapierrel and Gogarten 2009), a host can rapidly expand its metabolic potential, thereby extending both its ecological versatility and responsiveness to environment change (McFall-Ngai et al. 2013).

7. Escherichia

Escherichia spp., a member of *Enterobacteriaceae*, are oxidase-negative catalase-positive straight rods that ferment lactose. *E. coli* is a natural and essential part of the bacterial flora in the gut of humans and animals. Most *E. coli* strains are nonpathogenic and reside harmlessly in the human colon. However, certain serotypes do play a role in diverse intestinal and extraintestinal diseases (Kaper et al. 2004). In a study of the enteric bacteria present in the feces of Australian mammals, Gordon and FitzGibbon reported that *E. coli* was the commonest species, being isolated from nearly half of the hosts studied (Gordon and FitzGibbon 1999). Since it is also widely present in the environment, *E. coli* could function as a mediator for gene flow between environmental

and clinical settings (Patyar et al. 2010). *E. coli* species is constituted by a high number of strains, which can be found in multiple and distinct environments: industrial, metal-contaminated coastal environments like strain SMS-3-5 (Kaper et al. 2004; Fricke et al. 2008), living as commensals in the human gut like *E. coli Se11* (Oshima et al. 2008) and *E. coli Se15* (Toh et al. 2010) or even being used in the laboratory, as *E. coli* K011, which is a popular ethanol-producing strain (Ohta et al. 1991; Hammami et al. 2007).

8. Environmental variables

8.1. Nutrient availability

Nutrients are the chemicals or elements utilized for bacterial growth, uptaken from the environment the organisms inhabits. The major elements a bacterium such as *E. coli* needs are C (carbon), H (hydrogen), O (oxygen), N (nitrogen) and P (phosphorus). Other elements also necessary to bacterial growth but in less amount: S (sulfur), K (potassium), Mg (magnesium), Fe (iron), Ca (calcium), and the trace elements, which are metallic elements also necessary for bacterial growth but usually present in untraceable quantities like Mn (manganese), Zn (zinc), Co (cobalt), Cu (copper), and Mo (molybdenum). In humans, the necessary nutrients for survival and healthy maintenance of the body are mainly carbohydrates, Saccharides, Fats (which provide C, O and H), proteins (C, H, O, N and sometimes S) and amino acids. Salts (Na; K; Cl, chlorine) but also other minerals like Ca, P, S, Cu, Mg, Fe, I, Fl, Zn, Co and Se are also essential elements in the human healthy diet. Bacterial nutritional requirements are this way guaranteed, and consequently nutrient availability is never recognized as a limitation factor for growth of enteric bacteria within the host's gut. All dietary compounds that escape digestion in the small intestine are potential substrates of the bacteria in the colon. The bacterial conversion of carbohydrates, proteins and nonnutritive compounds such as poliphenolic substances leads to the formation of a large number of compounds that may have beneficial or adverse effects on human health (Leclerc et al. 2001).

8.2. Temperature

Temperature is one of the environmental variables that most influences bacterial survival and behavior. Each organism has a very precise range of temperatures in which it can survive, and within this range there is an optimum temperature, in which the growth of the organism is maximized and reproductive rate is increased. Comparing the enteric with the outer environment, it is expected in the enteric environment an enrichment of mesophile bacteria, with optimal growth temperature near 37ºC, in opposition to the optimal growth temperature near 20ºC of psychrophiles, usually enriched in the outer environments. Temperature has a great influence on organism's lifestyle, as it has been reported as a factor influencing the ecological, physiological and genomic properties of bacterial organisms (Zheng and Wu 2010), for example, in the population-level it has been shown to influence functional shifts in microbial communities, at the cellular-level, virulence functions and at the molecular-level it has been shown to influence codon usage and nucleotide content.

Psychrophiles have enzymes that are capable of performing their role more efficiently at lower temperatures. These organisms have in their cell membrane a high concentration of unsaturated fatty acids, which allows membrane fluidity at lower temperatures. On the other hand, enzymes from thermophile organisms are much more stable to high temperatures and the lipids they have in their cell membranes are richer in saturated fatty acids, allowing for fluidity in higher temperatures.

8.3. pH

pH stands for the measure of the acidity of an aqueous solution. Like with temperature, every organism has an optimal pH value, within a range of pH in which it can survive. Acid survival is defined as the ability of neutrophilic bacteria to survive at pH levels too acidic to permit growth. In the human gut, enteric bacteria experience variable oxygen and pH levels (Laing et al. 2011). The human intraluminal pH is rapidly exchanged from highly acid in the stomach to about pH 6 in the duodenum. The pH gradually increases in the small intestine from pH 6 to about pH 7.4 in the terminal ileum. The pH drops to 5.7 in the caecum, but again gradually increases, reaching pH 6.7 in the rectum. The

microbial intracellular pH is usually close to neutrality, maybe because the cell membrane is relatively impermeable to the entrance of protons. Exclusion of oxygen has been proposed to enhance acid survival, because anaerobic growth increases expression of acid stress mechanisms (McFall-Ngai et al. 2013). Most microorganisms inhabiting the human gastrointestinal tract are found in the duodenum, where pH is close to neutrality. Non the matter, anaerobic cultures of E. coli K-12 W3110 strains have been shown to survive in M63 minimal medium pH 2,5 (Laing et al. 2011). pH has the same type of influence on microbial growth that temperature does, as acidity levels too influence all organism's proteins and may affect temporarily or permanently the protein configuration and consequently its function.

8.4. Water activity (a_w)

The Aw measurement was developed to account for the intensity with which water associates with various non-aqueous constituents and solids. Water is the main compound of living organisms. Most microorganisms are only able to survive at a_w 0.98 or higher but there are some organisms which are capable of living in low a_w conditions (high concentrations of salt or sugar or even in conditions of dehydration). Dealing with enteric bacteria, the a_w factor is not considered a limiting factor.

9. Objectives

The main objective of this work is to find the genes responsible for the survival of microorganisms in the gut of homeothermic hosts, considering its physical and chemical properties:

- Do different *taxa* have different survival and adaptive strategies?
- Which features are shared among all the groups?

Considering the fact that our own species thrives within this ecosystem, it is necessary and pro-species survival to answer this question of how do things really work and relate with each other, because that would bring us closer to a healthier maintenance of our own home. Focusing on the biological sciences, McFall-Ngai et al. (2013) defends that applying metacommunity and network analyses to animal-bacterial interactions will be

essential for the design of effective strategies for managing ecosystems in the face of environmental perturbations, such as pollution, invasive species, and global climate. Furthermore, McFall-Ngai et al. 2013 states that whether an ecosystem is defined as a single animal or the planet's biosphere, the goal must be to apply an understanding of the relationships between all organisms within their environment, highlighting th importance of understanding relations between microbes and other organisms in order to be able to predict and manipulate microbial community structure and activity so as to promote ecosystem health (McFall-Ngai et al. 2013). Our work exploits the possibility of using only *in silico* analysis to perform many kinds of studies, under the perspective that it is possible to acknowledge genomes as groups of genes, which can be summed or intersected, without compromising conclusions about each gene's function or location.

Materials and Methods

1. Selection of the Working Model

Escherichia spp. was the selected taxon to carry out the data mining procedures within this work, as organisms from this division are present in multiple environments and consequently subjected to different selective pressures. This is a perfect scenario for application of the model presented in this work, aiming to compare closely-related genomes with variant environmental backgrounds.

2. Data Mining

The data mining was performed using the software EDGAR 1.2 (2009 @ Cebitec Bielefield University), namely Venn diagram and Geneset Calculation tools. Results retrieved from this first analysis were used as input for an extensive Blast analysis, performed using the Basic Alignment Search Tool.

3. Application of the model

Figure 1 presents the general workflow's diagram applied in this study. After selecting Escherichia spp. organisms as the primary study object, five genomes from strains with 3 distinct environmental backgrounds were used as input for a double-stepped EDGAR analysis aiming to isolate the genes exclusively present in genomes from organisms able to colonize the gut of homeothermic hosts. The resulting set of genes was subjected to a comprehensive BLAST analysis, performed with the goal of understanding which of those genes were in fact exclusive from organisms able to survive within the gastrointestinal tract of homeothermic hosts when compared to the NCBI nucleotide database. Identity and niche-specific thresholds were applied sequentially: the first one had the objective of excluding from each list of organisms retrieved by BLAST the organisms presenting low similarity homologs in relation to the query sequence, while the latter had the objective of excluding all query genes

presenting significantly similar homolog sequences in genomes of organisms which have never been reported as able to survive in the gut of homeothermic hosts.

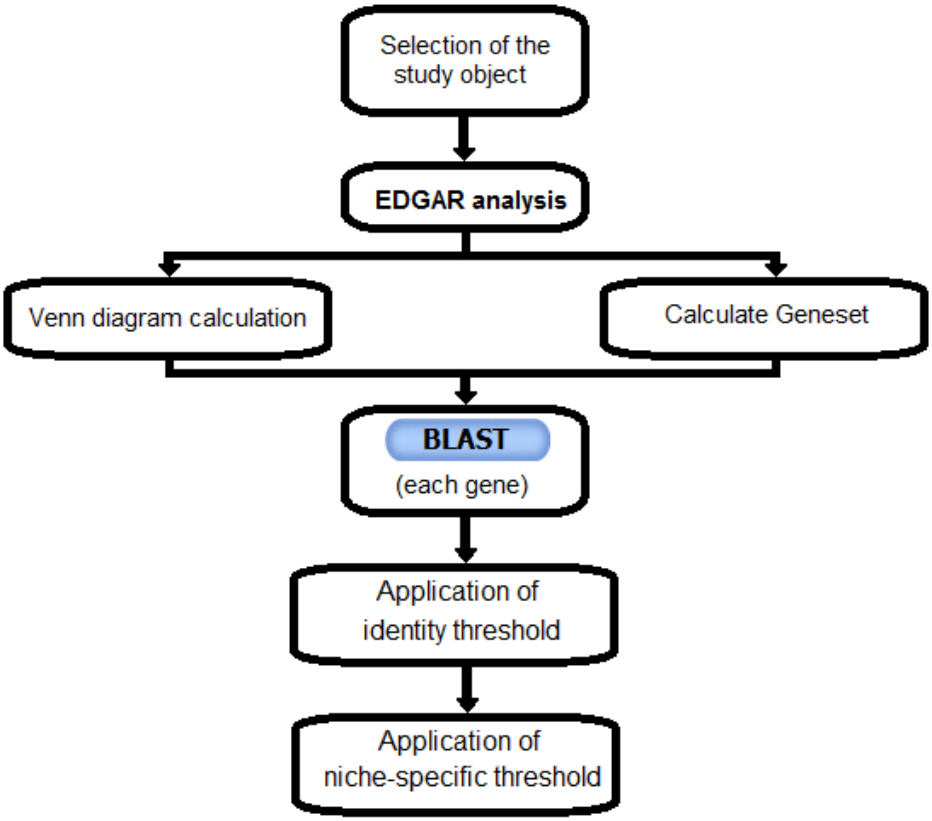

Figure 1- Study workflow diagram, resuming the main steps of the experiment.

EDGAR 1.2 was used to create a Venn diagram for five genomes of *E. coli* strains: SE11, SE15, SMS-3-5 and two K011. This tool allows the calculation of common gene pools (table 1).

Table 1 – *Escherichia coli* strains used to determine the gene set shared by two human gut commensal strains, SE11 and SE15, and absent from the genomes of three strains from the outer-gut environment, SMS-3-5, K011 and K011FL.

EDGAR's database strain	NBCI identification	Capability of Surviving within the gut of homeothermic hosts	Use in the Calculation of the Gene Set
Escherichia coli SE11	NC_011415	Yes	Included
Escherichia coli SE15	NC_013654		
Escherichia coli SMS-3-5	NC_010498	No	Excluded
Escherichia coli K011	NC_016902		
Escherichia coli K011FL	NC_017660		

The 77 genes included in the gene pool common to *E. coli* strains SE11 and SE15 and absent from SMS-3-5 and K011 were then individually used as query to search for somewhat similar sequences (blastn) in the nucleotide collection (nr/nt) database. This resulted in 154 lists of organisms with genomes that contain the same or a similar genetic sequence for each CDS (Table S1 to Table S154). Two thresholds were applied over the 154 CDS blastn analysis: the first threshold excluded from each list all the organisms whose e-value was at least 30 times lower than the e-value of the organism located immediately above it, i.e. organisms whose genomes presented no homologs for the query sequence; the second threshold excluded all CDSs related to organisms not reported as capable of surviving within the targeted niche: the gut of homeothermic hosts. From the total list of organisms left in the study, representatives for each species were selected based on NCBI calculations (NCBI's genomic database presents some strains as species' representative elements). These representative organisms were mainly chosen to represent the species lifestyle and some of its genomic features. Ultimately, each CDS was associated to the number of strains' genomes from the same species it exists in (Table 2 from Results and Discussion section).

Results and Discussion

1. EDGAR analysis disclosed the genes shared by commensal *E. coli* strains Se11 and SE15

The objective of this work was to identify genes particularly enriched or persistently present in the genomes of bacteria adapted to the gut of homeothermic hosts in order to disclose the genetic footprint of the corresponding bacterial ecotype. Starting from a primary comparison of the genomes of *E. coli* SE11, SE15, SMS-3-5 and two K011 strains using EDGAR 1.2 (2009 @ Cebitec Bielefeld University) (Blom et al. 2009), it was possible to retrieve the genes shared by the human gut isolated strains SE11 (Oshima et al. 2008) and SE15 (Toh et al. 2010) and absent from environmental and laboratory strains SMS-3-5 (Fricke et al. 2008) and K011 (Ohta et al. 1991; Dien et al. 1998). EDGAR, a bioinformatics platform for comparative genomics, has been designed to support the high throughput comparison of related genomes. EDGAR provides a database of high throughput comparison-based projects, each dedicated to the comparison of the genomes of each genus within the NCBI's database with more than three sequenced strains. In this study, EDGAR was employed to compare the genomes of *Escherichia coli* strains from different environmental backgrounds. EDGAR provides a Venn diagram calculation tool which allows the comparison of genomes belonging to the same genera, presenting results in a clear, easy-to-read graphic representation, and retrieving information on the number of specific, shared and core genes in the universe exclusively composed by the strains included in the calculation. The Venn diagram calculation using *E. coli* strains SE11 and SE15, both human commensal strains, SMS-3-5, isolated from metal-contaminated environmental waters, and two K011, both used in the laboratory for ethanol production, allowed to infer the amount of genes shared by the gut commensal strains and not present in the other strains. To theoretically intersect more than 5 genomes is possible but its visualization in the form of a Venn diagram would be rather confusing (Blom et al. 2009). The Venn diagram calculated using five genomes from *E. coli* strains disclosed 77 ortholog genes shared by the two strains able to survive within the gut of homeothermic hosts,

particularly in the human gastrointestinal tract in a commensal manner, SE11 and SE15 and absent from the genomes of the three strains isolated from outer environments, SMS-3-5 and both K011 (Figure 2).

2. The Baas-Becking Hypothesis

Escherichia spp. has been isolated from a multitude of environments, showing that these bacteria are constitutive part of numerous niches. Figure 3 presents the phylogenetic dendogram (from NCBI, calculated based on genomic blast) of *Escherichia* spp. and information on habitat is provided by the colored background. In a first rough analysis it does not seem to exist a strong correlation between relationship and ability to survive in a determined habitat.

The Baas-Becking hypothesis was proposed in 1934 and it presents the idea of "Everything is everywhere, but, environment selects". Microbial biogeographic studies show that the bacterial-taxa geographic distribution is not homogeneous and, at the same time, bacterial organisms' distribution is recognized as ubiquous (Fierer 2008). In the genomes of each strain are included the genes responsible for the adaptation of that strain to the environmental conditions of the habitats it is able to colonize. In order to cope with competitors, available nutrients or a determined range of temperatures and oxygen availability, each species has undergone a specific evolutionary pathway, which through time and generations has determined which genes were not to be maintained in each population. Putatively, every organism has the ability to colonize each habitat, acting as an incoming source of genes towards the habitat's gene pool. Yet, if their fitness level is low, the organisms may not be able to persist or successfully reproduce, thus making it more difficult for their specific genes to be maintained or enriched in the population. Fitter variants, capable of a successful colonization and reproduction, by maintaining their genomes in the habitat a longer period of time and even in a larger quantity, become more prominent gene donors to the habitat. Strains from the same habitat share the phenotypes that allow for their survival and adaptation

– yet, this can happen for two reasons: sharing genes (due to VGT[4] and HGT events (Ochman et al. 2000; Kurokawa et al. 2007; Harrison and Brockhurst 2012)) or having analogous genes providing the same metabolic functions. All things considered, it is expected that the results of the presented study underline and highlight the enrichment in the microbiome of genes responsible for the metabolic functions needed for the survival of bacteria in a targeted environment, due to the contribution of these genes to the fitness of the organisms carrying them. It is also expected that this enrichment lies on several variants of the majority of genera inhabiting the habitat, as *Escherichia* spp., *Klebsiella* spp. or even *Salmonella* spp., due to HGT events, prompted by the high concentration of cells in a rich and dynamic environment. Due to the environmental conditions of the gut of homeothermic hosts, this is considered a complex environment – the lack of extreme physical and chemical variables allows for the successful colonization and survival of multiple taxa within the environment, which, at the same time, constitutes a challenge to other taxa, by increasing the competitiveness for the same resources.

[4] Vertical Gene Transfer

3. The Venn Diagram

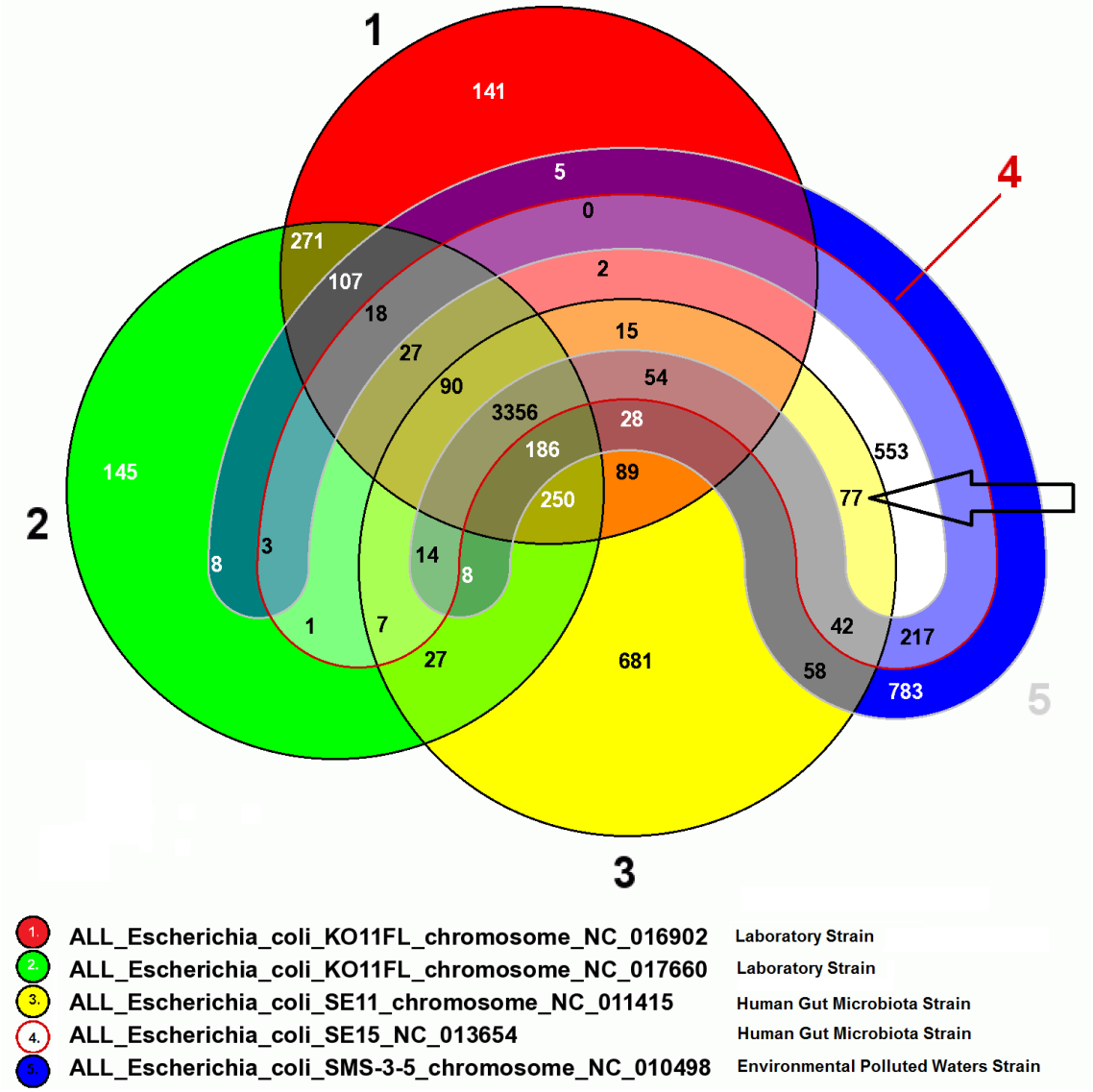

Figure 2 - Venn diagram calculated using *E. coli* strains K011, K011, SE11, SE15 and SMS-3-5, accession numbers respectively, NC_016902, NC_017660, NC_011415, NC_013654, NC_010498.

In figure 2 it is represented the Venn diagram calculated with the goal to find how many genes are shared between SE11 and SE15 and absent from the genomes of strains SMS-3-5 and K011. The higher number of shared genes (3356 genes) corresponds to the area intersected by all strains, representing the core genome for this group of strains. As only *Escherichia coli* strains were included in the study, this core genome is presumed to account for housekeeping genes, responsible for basic cell maintenance and *Escherichia coli* specific genes.

Table 1 – Resume of lowest number of shared genes when intersecting *Escherichia coli* strains K011FL, SE15, SE11 and SMS-3-5.

The intersection of *Escherichia coli* strains:					Reported # shared genes:
K011FL NC_016902	**with**	SE15 NC_013654			2
		SE15 NC_013654	**and**	SMS-3-5 NC_010498	0
		SMS-3-5 NC_010498			5
K011FL NC_017660	**with**	SE15 NC_013654			1
			and	SE11 NC_011415	7
				SMS-3-5 NC_010498	3
		SMS-3-5 NC_010498			8
		SMS-3-5 NC_010498	**and**	SE11 NC_011415	8

A rather interesting intersection is the one involving *E. coli* K011FL (NC_016902, corresponding to area number 1), *E. coli* SE15 (corresponding to area number 4) and *E. coli* SMS-3-5 (corresponding to area number 5), which does not present any gene. This is curious in the matter that each of the three strains came from a different environmental background, correspondingly the laboratory, the human gut and environmental polluted waters, which could explain the lack of shared genes between these strains. In table 1 are highlighted the intersections of genes which revealed less than ten shared genes.

Table 2 - Sum of genes in the intersection of presented strains. The darker shadow represents the total number of genes in each strain (diagonal line).

	K011FL NC_016902	K011FL NC_017660	SE11 NC_011415	SE15 NC_013654	SMS-3-5 NC_010498
K011FL NC_016902	4639				
K011FL NC_017660	4305	4508			
SE11 NC_011415	4068	3958	4982		
SE15 NC_013654	3562	3516	3655	4476	
SMS-3-5 NC_010498	3754	3700	3746	3704	4887

Based on the calculated Venn diagram, table 2 was calculated: it presents the amount of genes shared by each pair of strains used in the EDGAR analysis. The higher number of shared genes (4305) corresponds to the intersection of both K011FL genomes, which could be related to the fact that these strains are more closely related to each other than any other pair of strains. In figure 3, the genetic proximity between both K011 strains is again clear, compared to pairwise relationships between other strains. In this set of 4305 shared genes are presumably included genes responsible for the proximity of the laboratory strains K011FL relative to both lineage (identity by descent) and colonization of niche (identity by descent and HGT). The lower number of shared genes (3516) is between SE15 and K011FL (NC_017660) in table 2, and this pairwise relationship corresponds to the higher genetic distance in phylogenetic tree from figure 3.

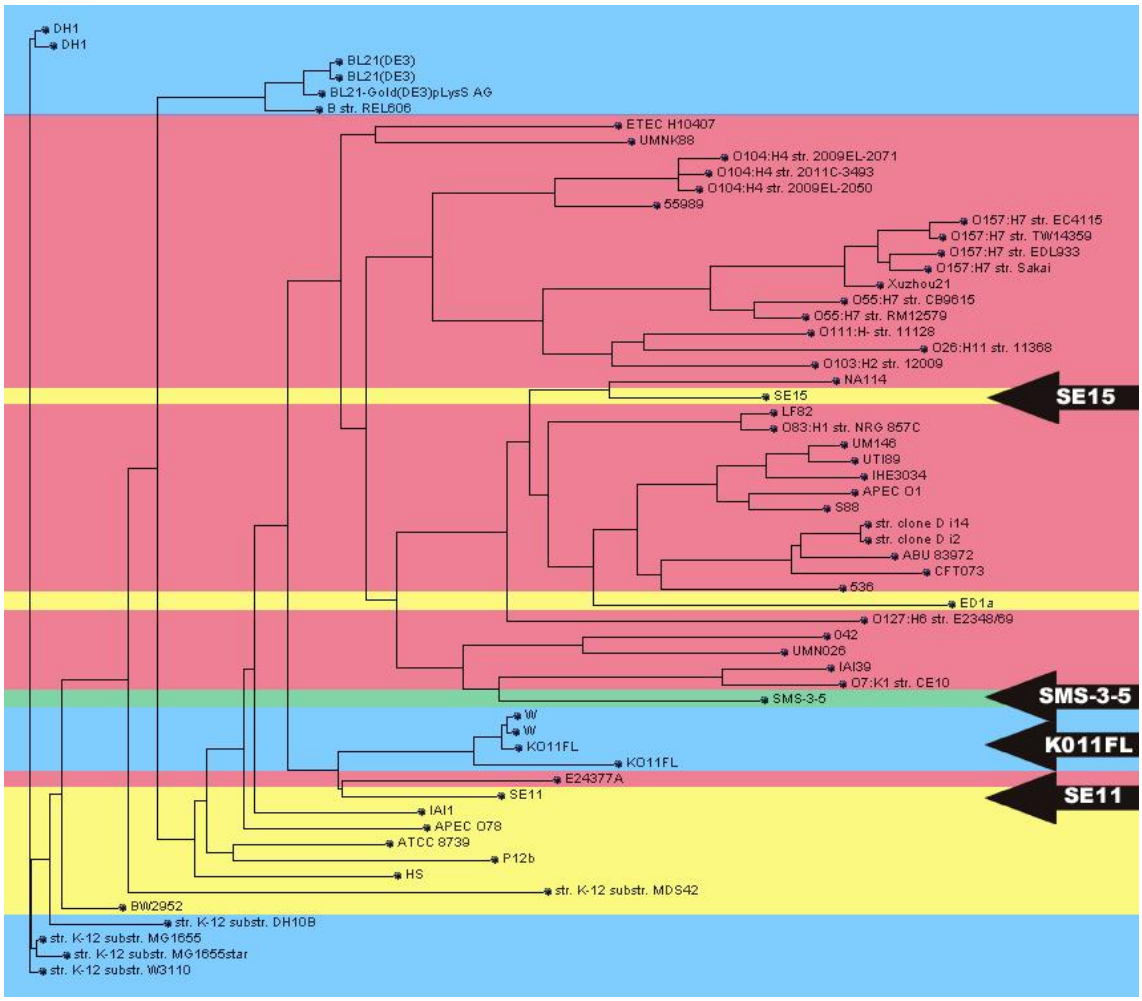

Figure 3 – Dendogram distance tree of *E. coli* strains, retrieved from NCBI, calculated using whole genome blast alignments. Distances are based on pairwise BLAST scores for each genome pair. Legend: Blue – laboratory strains; Yellow – Commensal to homeothermic hosts; Green – environmental; Pink – Strict / potential pathogen to homeothermic hosts. The arrows highlight the position of strains SE15, SE11, SMS-3-5 and both K011FL.

As bacterial genomes lose and gain genes, a continuous modulation of bacterial organisms' fitness contributes to the fixation of some variants through a selection process which includes genetic drift and positive selection of fitter variants. Bacterial genomes present a varied distribution of mobile genetic elements and metabolic islands[5], which together with the reported gene acquisition via HGT[6] show a high level of genomic plasticity[7]. Furthermore, the gut of homeothermic hosts, particularly the human gastrointestinal tract, has been reported as a "hot spot" for HGT between

[5] Clusters of genes with defined and specific metabolic functions
[6] Horizontal Gene Transfer
[7] Re-arrangement of genomic regions between species

microbes, fact emphasized by the abundance of mobile genetic elements found in the human intestinal gene pool[8] (Kurokawa et al. 2007). The presence or absence of a gene in the genome of an organism, or even its modification, is a major factor determining the potential metabolic capabilities of the organism in its habitat (Altermann 2012). Considering the genomic plasticity of bacterial genomes, particularly demonstrated within the human gut's microbiota, this work first proposes that the set of genes shared by two commensal *E. coli* strains and absent from three strains not related to this type of environment (same species, different ecotype) includes genes related to the survival of bacterial organisms in this ecotype, providing the means to further calculate the bacterial ecotype genetic footprint. By comparing the genomes of strains from different environmental backgrounds using Venn diagrams, one gets a clear visualization of common gene pools, pan-genomes and singletons. SE11 and SE15 were used in this context to represent the organisms able to survive and colonize the gut of homeothermic hosts, while SMS-3-5 and both K011 strains were used to represent the organisms inhabiting outer environments, more specifically the outside of the gut of homeothermic hosts. Yet, to use exclusively *Escherichia coli* strains' genomes makes the results biased. Accordingly, the genes found in the gene set shared by the strains able to survive in the targeted environment are expected to be classifiable in roughly two groups: (i) genes related to the metabolic functions related to the survival of bacterial organisms within the gut of homeothermic animals, which are expected to be present in the genomes of non-*E. coli* strains also capable of surviving within the gut of homeothermic animals and (ii) *Escherichia coli* specific genes maintained in the genomes of independent strains SE11 and SE15 and absent from the genomes of strains SMS-3-5 and K011s by genetic drift (not by positive selection). The genes related to the metabolic functions related to the successful establishment of bacterial organisms within the gut of homeothermic animals, in this context, are the genes (i) shared by *E. coli* strains SE11 and SE15 and absent from strains SMS-3-5 and K011s, and (ii) present in the genomes of other non-*Escherichia* strains which also successfully colonize the gut of homeothermic animals.

To gain further insight into which genes are associated with the survival of bacterial organisms within the gut of homeothermic animals, the DNA sequences of each one of these 77 ortholog pairs of genes were obtained using EDGAR's tool calculate genesets

[8] Set of genes persistently present in a determined niche, within the genomes of inhabiting strains.

and 154 FASTA-formatted CDSs were retrieved. Each one of the 154 CDSs was used as query in BLASTn to search for somewhat similar sequences against the nucleotide database (Table S1 to S154). This procedure revealed which strains had each one of these genes in their genomes, organized in lists by similarity to the query sequence, which allowed the sequential application of two exclusive thresholds:

1. The first organisms within the resulting list with an e-value 30 times lower than the organism immediately above it was excluded, along with all organisms below it (identification threshold);
2. CDSs present in organisms known to be unrelated with gut colonization, were excluded from further analysis (niche-specific threshold).

The application of the first threshold, the identity threshold, is an attempt to remove from the work-on sample the organisms that have in their genomes CDSs that greatly differ from the query sequence, and therefore they are not considered homologous genes. The e-value describes the significance of the match of each retrieved sequence to the query sequence: it is the number of hits one can expect to see by chance when searching a database of a particular size. The lower the e-value, the more “significant” the match is, which means that as the e-value slowly increases, the significance of the match decreases - the sequences are less and less similar to the query sequence. The e-value can be used as a significance threshold – if the statistical significance ascribed to a match is greater than the expect threshold, the match will not be reported. Yet, in this work there was no expect threshold defined *a priori*. The threshold applied is based on the e-value’s difference from one strain to the next: if the e-value showed a radical increase (30 times or higher difference), the organisms below in the list were excluded, because the similarity of the retrieved sequence to the query sequence is not considered “significant enough” for that organism to be identifiable by the query sequence. In this context, this radical increase of the e-value was usually associated to a significantly lower quality alignment, due to the smaller length of the retrieved sequence.

The application of the second threshold, the niche-specific threshold, was performed as an attempt to remove from the work-on sample every CDS that is not related to the survival within the targeted environment. Each CDS was associated with the list of organisms in which genomes it exists, organized by e-value. After applying the

identification threshold, each list contained only organisms that would be identified by the presence of each CDS in their genomes. Whole lists of organisms associated to their CDSs were excluded from the work-on sample because they included organisms not related to the survival within the niche specified. The presence of a significantly similar sequence to the query sequence in the genome of an organism not found in the gut of homeothermic animals leads to its exclusion from the study because this reveals its lack of niche specificity needed for a genetic footprint pretending to characterize a bacterial ecotype. This process led to the exclusion of 83 CDSs (Table S1 to S154 of Supplementary Material).

The set of genes left in the work for further analysis (52 chromosomal and 19 plasmidic sequences) includes genes putatively related to the survival of bacterial organisms in the gut of homeothermic animals and *Escherichia coli* or *Escherichia* spp. specific genes. There was no threshold applied to exclude genes whose presence was only validated in genomes from *Escherichia* spp. strains, yet they could have been excluded from the study: if they are present only in strains from the same genera and colonizing the same kind of ecotype (homeothermic host's guts), they are most probably not related to the niche adaptation.

Table 3 - Description of some strains used as representatives of species retrieved from the application of both thresholds. C.g. stands for complete genome. The column named 'Habitat' holds information on the distribution of the organisms according to their habitat, based in the classification parameters present in the Genome Project for each genome available in NCBI: 1, terrestrial; 2, aquatic; 3, multiple; 4, host-associated; 5, specialized. The column named 'Optimal Growth Temperature' holds information on the distribution of the organisms according to their optimal growth temperature, based on the classification parameters of the Genome Project for each genome available in NCBI: 0, unknown/undefined; 1, psychrophilic; 2, mesophilic; 3, thermophilic; 4, hyperthermophilic. The column named 'Relationship With the Host' holds information on the distribution of the organisms according to their relationship with the host, based on the information in the Genome Project for each genome available in NCBI, as well as in the papers describing the sequencing of each specific genome: 1, no association; 2a, strict symbiosis/commensalism with animals; 2b, strict symbiosis/commensalism with plants; 3a, facultative symbiosis/commensalism with animals; 3b, facultative symbiosis/commensalism with plants; 4a, strict pathogen in animals; 4b, strict pathogen in plants; 5a, facultative pathogen in animals; 5b, facultative pathogen in plants; 5c, facultative pathogen in bacteria; 6, bacterial commensalism. Information on genome size, G+C content and Leclerc's Group is also presented.

Representative Strains	Habitat	Optimal Growth Temperature	Relationship With the Host	Genome Size (Mb)	%G/C	Leclerc's Group
C. koseri ATCC BAA-895	3	2	3a	4,7	54	2
C. rodentium ICC168	4	3	3a	5,4	55	?
E. aerogenes KCTC 2190	4	3	5a	5,3	55	2
E. asburiae LF7a	4	3	5a	5	54	?
E. cloacae subsp. *cloacae* ATCC 13047	4	3	5a	5,6	55	2
E. coli IAI39 chromosome	4	3	5a	5.13	50.6	1
E. coli O104:H4 str. 2011C-3493	4	3	5a	5.44	50.6	1
E. coli O157:H7 str. Sakai DNA	4	3	5a	5,6	50	1
E. coli O83:H1 str. NRG 857C	4	3	5a	4,9	50.7	1
E. *coli* UMN026	4	3	5a	5,4	50.6	1
E. coli str. K-12 substr. MG1655	4	3	5a	4.64	50.8	1
E. fergusonii ATCC 35469	4	3	5a	4,6	50	?
K. oxytoca KCTC 1686	4	3	5a	6	56	2
K. pneumoniae KCTC 2242	3	2	3a	5,5	57	2
K. pneumoniae subsp. *pneumoniae* HS11286	3	2	3a	5,7	57	2
K. pneumoniae subsp. *pneumoniae* MGH 78578	3	2	3a	5,7	57	2
K. variicola At-22	3	2	3ab	5,5	58	?
S. bongori NCTC 12419, culture collection SGSC SARC11	4	3	4a	4,5	51	?
S. enterica srv. *Typhi* (S. typhi) strain CT18	4	3	4a	5,1	52	?
S. enterica subsp. *enterica* srv. Paratyphi A str. ATCC 9150	4	3	4a	4,6	52	?
S. enterica subsp. *enterica* srv. Typhimurium str. 14028S	4	3	4a	5	52.2	?
S. enterica subsp. *enterica* srv. Typhimurium str. LT2	4	3	4a	5	52	?
S. plymuthica AS9	3	2	3b	5.44	56	?
S. proteamaculans 568	3	3	4a	5,5	55	?
S. marcescens WW4	3	2/3	3b/ 4b	5.24	59.6	?
Y. intermedia strain 6270	3	2/3	3a	4.71	47.4	3
Y. pseudotuberculosis YPIII	3	2/3	3a	4,7	48	3
S. boydii Sb227	5	3	4a	4,7	51	?
S. dysenteriae Sd197	5	3	4a	4,6	51	?
S. flexneri 2a str. 301	5	3	4a	4,8	51	?
S. sonnei Ss046	5	3	4a	5,1	51	?

From the total set of organisms present in the work-on sample, representative strains were chosen based on NCBI data (table S155 of Supplementary Material). These strains are used here to represent each species. CDS present in their genomes – 48 chromosomal and 19 plasmidic genes - remained in the study for further analysis. Representative strains are characterized by some factors related to the capability of

surviving within the gut of homeothermic hosts (table 3). None of the five strains used in the EDGAR comparative analysis is a representative for *Escherichia* spp. In this part of the work the goal was to relate the presence of each gene in the genomes of the representative strains to the persistency degree of the gene in each species (Table 4).

Table 4.1 - Quantitative measurement of the persistency of each plasmid CDS within the genome of each species considering the corresponding representative strains - numbers account for the number of representative strains of each species presenting the CDS in their genomes and points corresponds to absence of genes. Legend: HP – Hypothetical protein; ptt. – Putative.

	Citrobacter spp.	*Enterobacter* spp.	*Escherichia* spp.	*Klelbsiella* spp.	*Salmonella* spp.	*Serratia* spp.	*Yersinia* spp.	*Shigella* spp.	TOTAL GENE	Function according to EDGAR
ECSE_P1-0020	.	.	1	.	.	.	.	1	2	colicin immunity protein
ECSF_P1-0085	.	.	1	.	.	.	.	1	2	HP
ECSE_P1-0021	.	.	1	.	.	.	.	1	2	HP
ECSF_P1-0084	.	.	1	.	.	.	.	1	2	HP
ECSE_P1-0023	.	.	2	.	.	.	.	1	3	HP
ECSE_P1-0046	.	.	3	.	.	.	.	1	4	HP
ECSF_P1-0134	.	.	3	.	.	.	.	1	4	HP
ECSE_P1-0048	.	.	1	.	.	.	.	1	2	HP
ECSF_P1-0136	.	.	2	.	.	.	.	.	2	HP
ECSE_P1-0049	.	.	3	.	.	.	.	1	4	HP
ECSF_P1-0137	.	.	3	.	.	.	.	1	4	HP
ECSE_P2-0012	.	.	3	2	1	.	.	.	6	Resolvase
ECSF_P1-0119	.	.	5	1	.	.	.	.	6	Resolvase
ECSE_P2-0042	.	.	4	.	.	.	.	.	4	HP
ECSF_P1-0150	.	.	4	.	.	.	.	.	4	HP
ECSE_P2-0044	.	.	3	.	.	.	.	.	3	plasmid SOS inhibition protein B
ECSF_P1-0002	.	.	3	.	.	.	.	.	3	regulator of SOS induction PsiB
ECSE_P2-0089	.	.	1	.	.	.	.	.	1	replication protein
ECSF_P1-0064	.	.	1	.	.	.	.	.	1	replication protein RepB

Table 4.2 - Quantitative measurement of the persistency of each chromosomal CDS within the genome of each species considering the corresponding representative strains - numbers account for the number of representative strains of each species presenting the CDS in their genomes and points corresponds to absence of genes. Legend: HP - Hypothetical protein; ptt. – Putative.

	Citrobacter spp.	*Enterobacter* spp.	*Escherichia* spp.	*Klelbsiella* spp.	*Salmonella* spp.	*Serratia* spp.	*Yersinia* spp.	*Shigella* spp.	TOTAL GENE	Function according to EDGAR
ECSE_0030	1	1	3	.	4	.	.	4	13	HP
ECSF_0034	.	.	1	.	.	.	.	1	2	HP
ECSE_0122	2	1	3	.	4	.	.	4	14	HP
ECSF_0135	2	1	3	.	4	.	.	4	14	HP
ECSE_0267	.	.	1	.	.	.	.	.	1	HP
ECSF_4239	.	.	1	2	.	.	.	.	3	HP
ECSE_0269	.	.	1	1	.	.	.	.	2	HP
ECSF_4241	.	.	1	1	.	.	.	.	2	ptt. acetyltransferase
ECSE_0327	1	.	.	.	.	.	.	.	1	ptt. autotransporter
ECSF_0284	1	.	.	.	.	.	.	.	1	HP
ECSE_0331	.	.	.	.	5	.	.	.	5	ptt. phage integrase
ECSF_0288	.	.	.	.	5	.	.	.	5	ptt. phage integrase
ECSE_0393	2	.	3	.	5	.	.	4	14	ptt. autotransporter
ECSF_0334	2	.	3	.	5	.	.	4	14	flagellar protein
ECSE_0562	.	.	1	.	.	.	.	.	1	HP
ECSF_1039	.	.	1	.	.	.	.	.	1	HP
ECSE_0566	.	.	2	.	.	.	.	.	2	HP
ECSF_1043	.	.	1	.	.	.	.	.	1	phage protein
ECSE_0567	.	2	3	.	.	.	.	1	6	phage exonuclease
ECSF_1044	.	2	2	.	1	.	.	1	6	phage exonuclease
ECSE_0569	.	.	.	.	.	.	.	.	0	phage host-nuclease inhibitor protein
ECSF_1046	.	.	2	.	.	.	.	.	2	phage host-nuclease inhibitor protein
ECSE_0572	.	1	1	1	1	.	.	.	4	phage transcriptional regulator
ECSE_0573	.	.	.	.	.	.	.	.	0	HP
ECSE_0574	.	.	.	.	.	.	.	.	0	HP
ECSE_0575	.	.	1	.	2	.	.	.	3	ptt. phage replication protein
ECSF_1052	.	.	1	.	2	.	.	.	3	ptt. phage replication protein
ECSE_0577	.	.	2	.	.	.	.	.	2	phage exclusion protein
ECSF_1054	.	.	2	.	.	.	.	.	2	phage exclusion protein
ECSE_0581	.	.	2	.	.	.	.	1	3	HP
ECSF_1058	.	.	2	.	.	.	.	1	3	phage protein

Table 4.2 (continuation) - Quantitative measurement of the persistency of each chromosomal CDS within the genome of each species considering the corresponding representative strains - numbers account for the number of representative strains of each species presenting the CDS in their genomes and points corresponds to absence of genes. Legend: HP - Hypothetical protein; ptt. – Putative.

	Citrobacter spp.	*Enterobacter* spp.	*Escherichia* spp.	*Klelbsiella* spp.	*Salmonella* spp.	*Serratia* spp.	*Yersinia* spp.	*Shigella* spp.	TOTAL GENE	Function according to EDGAR
ECSE_0584	.	.	2	.	.	.	.	1	3	endodeoxyribonuclease RUS
ECSF_1061	.	.	2	.	.	.	.	.	2	phage endodeoxyribonuclease
ECSE_0587	2	2	1	.	4	4	.	.	13	ptt. outer membrane porin protein
ECSF_1063	2	2	2	.	4	2	2	.	14	ptt. outer membrane porin protein
ECSE_0706	.	.	2	.	5	.	.	3	10	ptt. phage major capsid protein
ECSF_0577	.	.	3	.	5	.	.	2	10	HP
ECSE_1657	.	.	2	.	2	.	.	.	4	ptt. phage major capsid protein
ECSF_1074	.	.	2	.	2	.	.	.	4	ptt. phage major capsid protein
ECSE_1659	.	.	2	.	2	.	.	.	4	ptt. phage capsid assembly protein
ECSF_1072	.	.	2	.	2	.	.	.	4	ptt. phage capsid assembly protein
ECSE_3291	2	.	2	.	1	.	.	2	7	HP
ECSF_2836	2	.	3	.	1	.	.	3	9	HP
ECSE_3531	1	.	2	.	5	.	.	3	11	HP
ECSF_3075	1	.	3	.	5	.	.	1	10	HP
ECSE_3818	2	1	2	1	5	.	.	2	13	HP
ECSF_3380	2	.	3	1	5	.	.	2	13	HP
ECSE_4376	2	.	2	.	5	.	.	1	10	truncated formate dehydrogenase H
ECSF_3959	2	.	3	.	5	.	.	2	12	truncated formate dehydrogenase H
ECSE_4678	1	.	2	.	5	.	.	3	11	HP
ECSF_4337	1	.	2	.	5	.	.	3	11	HP

In tables 4.1 and 4.2 we can observe some interesting patters, which may be informative. Representative strains of each species were used to grossly understand the distribution of genes among different lineages sharing their habitats.

There is not a single gene shared by all species, which suggests that a genetic footprint will not rely only on the distribution of single gene throughout all species considered but yet will rely on the presence of determined set of genes, which presence will confirm the existence of the bacterial ecotype targeted in the first place. Genes considered to be good targets for the calculation of this genetic footprint are the

ones present in at least two genera and so, they are putatively related to the metabolic functions responsible for the adaptation of organisms to this specific niche. Genes present in one sole genus (20 genes, 7 plasmidic and 13 chromosomal) were considered not fit for the design of the genetic footprint of a bacterial ecotype because their function will be more probably related to the specificities of the organisms which make them part of a determined taxon, than to the capability of the organism to survive in a determined habitat.

Several genes were found to be exclusively present in genomes of both genera *Escherichia* and *Shigella*, namely plasmid genes ECSE_p1-0020, ECSF_p1-0085, ECSE_p1-0021, ECSF_p1-0084, ECSE_p1-0023, ECSE_p1-0046, ECSF_p1-0134, ECSE_p1-0048, ECSE_p1-0049, ECSF_p1-0137 and chromosomal genes ECSF_0034, ECSE_0581, ECSF_1058 and ECSE_0584. These genes were also considered not to be useful to calculate the genetic footprint of the bacterial ecotype inhabiting the gut of homeothermic hosts, mainly because the controversy about the relatedness level of organisms from these two genera rises up questions about the reason these genes were maintained in the genomes.

Serratia and *Klebsiella* genera include organisms reported in NCBI as capable of inhabiting outside the gut of homeothermic hosts, namely *Klebsiella variicola* At-22 and *Serratia plymuthica* AS9, used in this work as representatives. Consequently, genes reported as present in these genera must be excluded from the work. The reason they are present in the study at this point is the human error associated with the application of the niche-specific threshold. Genes excluded by this reason (lack of specificity to the targeted environment) were ECSF_p1-0119, ECSE_p2-0012, ECSF_4239, ECSE_0269, ECSF_4241, ECSE_0572, ECSE_3818, ECSF_3380, ECSE_0587 and ECSF_1063. This specific human error also highlights the readiness to determine the quality of the previous application of the thresholds: all these genes were supposed to have been excluded during the application of the second threshold, the niche-specific threshold, which has as main purpose to exclude from the gene set the genes that are not exclusively present in organisms adapted to the survival within the gut of homeothermic hosts, but it has not. None the less, one can observe how easy it is to detect and purify the gene set after is has been reduced to a “workable” number of shared genes.

Bacterial-host interactions have been demonstrated as an important matter of concern, due to the consequences both participants suffer – the host microbiota is a dynamic entity which changes through time, influencing the host and molding its characteristics. The bacterial community within a host changes in response to immigration of new species, host immune system pressures and selection by phages.

Koskella, B. et al (2012) demonstrated that the microbiota on a vegetable host's leaves altered throughout time gaining resistance to phages from the past. There is a dynamic relationship between phages and bacteria – phages exert pressure on the bacterial populations, eliminating the individuals lacking resistance, and molding this bacterial population's genomes in a way to maintain in the gene pool genes responsible for the resistance. The phage present in an environment will potentially act in all bacterial taxa without resistance, making the response (the evolution of resistance to this phage) a community or population event – one phage is able of select against multiple bacteria species, while those respond together, rather than individually. The enrichment of these genes is expected to be ubiquitous among the species inhabiting the same environment, if they are exposed to the same environmental pressures, phage-selection included.

Table 5 - Table resuming the 23 genes revealed as appropriate to design the genetic footprint of a bacterial ecotype adapted to the gut of homeothermic hosts.

ECSE_0030	HP
ECSE_0122	HP
ECSF_0135	HP
ECSF_0577	HP
ECSE_3291	HP
ECSF_2836	HP
ECSE_3531	HP
ECSF_3075	HP
ECSE_4678	HP
ECSF_4337	HP
ECSE_0567	phage exonuclease
ECSF_1044	phage exonuclease
ECSE_0575	putative phage replication protein
ECSF_1052	putative phage replication protein
ECSE_0706	putative phage major capsid protein
ECSE_1657	putative phage major capsid protein
ECSF_1074	putative phage major capsid protein
ECSE_1659	putative phage capsid assembly protein
ECSF_1072	putative phage capsid assembly protein
ECSE_0393	putative autotransporter
ECSF_0334	flagellar protein
ECSE_4376	truncated formate dehydrogenase H
ECSF_3959	truncated formate dehydrogenase H

This work revealed 23 genes fit for the calculation of the genetic footprint of the targeted bacterial ecotype. Ten of those genes are reported as responsible for the production of hypothetical proteins. A more thorough biochemical study on them could provide more clues about their function, yet, their exclusive presence in multiple species of bacteria adapted to the gut of homeothermic hosts leads to the conclusion that they must be related to the metabolic functions allowing for their successful colonization and adaptation.

Nine genes are related to the production of phage proteins – phage nuclease, phage transcription regulator and a phage protein, and these are present in multiple species inhabiting the targeted environment. Phages have been demonstrated as one of the most important selection factors acting on bacterial populations (time shift studies in

the laboratory showed consequent reactions of both bacteria and phages to the behavior of phages and bacteria respectively, and these conclusions have been extended to vegetal host's microbiota by the work of Koskella, B. et al. (2012). Furthermore, horizontal gene transfer events are often mediated by conjugative phages, accelerating the adaptation of the organisms to the habitat, aiding in the molding of their genomes (Fierer 2008).

The three remaining genes are reported as responsible for the production of (i) putative autotransporters, which function is confirmed by the work of Oshima, K. et al. (2008) (ii) flagellar proteins and (iii) truncated formate dehydrogenase H, which is a protein related to survival of organisms under anaerobic conditions, fulfilling all the requirements to become part of the genomic signature specific of the bacterial ecotype inhabiting the gut of homeothermic hosts.

Future Perspectives

This study started with a comprehensive genomic analysis aiming to target genes specifically present or enriched in genomes of bacterial organisms able to successfully colonize the gut of homeothermic hosts, namely the human. The results include genes present in at least 2 genera of bacteria able to inhabit in the gut of homeothermic hosts, suggesting the importance of these genes in the adaptation processes and evolutionary pathways this bacterial ecotype went through throughout time in this specific habitat. Interestingly, most of these genes code for the production of hypothetical proteins, or for phage-related proteins, namely phage nucleases and phage transcription regulators. To discern the importance of these putative gut-specific genes further studies are needed, particularly addressing the following topics:

Statistical analysis to determine gene persistence and enrichment for gut-specific genes

A statistical analysis of presence and persistence of genes in the genomes of organisms able to survive within the gut of homeothermic hosts would be useful, as it would provide information on the significance of the presented results, and better define enrichment levels of the targeted genes. If these genes definitely confer advantages for organisms in this determined environment, organisms holding them in their genomes will be more and more represented in each generation, while other lineages, competing with the first, will perish or be reduced to a small number of organisms, leading to the disappearance or low-level representation of their genetic patrimony. It would be necessary to analyze the significance of the enrichment levels of each gene, to understand if this enrichment is not just due to chance, because if so, this would disprove the gene as putatively responsible for the adaptation to this habitat. Also, there is the need to calculate the distance (genetic linkage) between each pair of genes found in the same genome, to guarantee their transmission from generation to generation is independent of the transmission of the physically nearest gene. This is important because every sample should be constituted of independent observations, and if two given genes are physically close enough, the transmission of one of the genes onto the next generation's gene pool may be due to the physical proximity to

another gene under some form of positive selection, instead of being due to the positive selection over the gene itself.

Gut-specific biomarkers

Results presented in this work should also provide a wider view on the biology of the bacterial community able to inhabit guts of homeothermic hosts, including the possibility of designing new environmental biomarkers which could be used as molecular indicators of fecal contamination in outer environments and food stuffs. Individually, each gene presented by this work as putatively related to the survival and adaptation of microbial organisms to the gut of homeothermic hosts, characterizes this type of environment, and therefore, theoretically, their presence is indicative of fecal contamination, as long as there is no notice of an HGT event introducing it in the genomes of lineages not related with this environment.

Disclosing the metabolic functions of hypothetical proteins

In a less urgent scenario, future studies are needed in order to disclose the metabolic functions of the reported hypothetical proteins, aiming for a better understanding of the targeted ecosystem. These proteins, by being coded exclusively by the genomes of organisms from multiple bacteria inhabiting the same habitat and not by genomes of closely-related organisms inhabiting other environments, are thought to be essential for the adaptation and survival of bacterial organisms under the physical, chemical and ecological conditions the environment presents. Accordingly, their role as providers of adaptive mechanisms conferring advantages to microbial organisms in this context must be exploited and understood, as part of a better global understanding of human and environmental health, bacterial metabolism and even the web-like distribution of information across genomes.

Evolutionary history of the putative gut-specific genes

Evolutionary pathways that led bacterial organisms to their actual distribution across habitats is a matter of interest. Reported phage proteins and stability of adaptation-related genes in the genomes of multiple genera of bacteria inhabiting the gut today, conduces to the idea of several episodes of introduction of phage genes in the genome of bacterial organisms happening a long time ago, a posterior maintenance of that

same gene within the bacterial lineage, and for some genes, a maintenance within the whole bacterial community. This, together with the analysis of position of genes in each reported genome would be interesting, as it would highlight horizontal gene transfer events which led to the presence of the same genes in vertically independent lineages, and consequent population-level enrichment of genes.

References

Altermann, E. (2012). "Tracing lifestyle adaptation in prokaryotic genomes." Frontiers in microbiology **3**: 48.

Backhed, F., H. Ding, T. Wang, L. V. Hooper, G. Y. Koh, A. Nagy, C. F. Semenkovich and J. I. Gordon (2004). "The gut microbiota as an environmental factor that regulates fat storage." Proceedings of the National Academy of Sciences of the United States of America **101**(44): 15718-15723.

Blaut, M. and T. Clavel (2007). "Metabolic diversity of the intestinal microbiota: Implications for health and disease." Journal of Nutrition **137**(3): 751S-755S.

Blom, J., S. P. Albaum, D. Doppmeier, A. Puhler, F. J. Vorholter, M. Zakrzewski and A. Goesmann (2009). "EDGAR: A software framework for the comparative analysis of prokaryotic genomes." Bmc Bioinformatics **10**.

Brosch, R., A. S. Pym, S. V. Gordon and S. T. Cole (2001). "The evolution of mycobacterial pathogenicity: clues from comparative genomics." Trends in Microbiology **9**(9): 452-458.

Cabral, J. P. S. (2010). "Water Microbiology. Bacterial Pathogens and Water." International Journal of Environmental Research and Public Health **7**(10): 3657-3703.

Collins, S. M., M. Surette and P. Bercik (2012). "The interplay between the intestinal microbiota and the brain." Nature Reviews Microbiology **10**(11): 735-742.

Crick, F. (1970). "Central Dogma of Molecular Biology." Nature **227**(5258): 561-&.

Dien, B. S., R. B. Hespell, H. A. Wyckoff and R. J. Bothast (1998). "Fermentation of hexose and pentose sugars using a novel ethanologenic *Escherichia coli* strain." Enzyme and Microbial Technology **23**(6): 366-371.

Ezenwa, V. O., N. M. Gerardo, D. W. Inouye, M. Medina and J. B. Xavier (2012). "Animal Behavior and the Microbiome." Science **338**(6104): 198-199.

Fricke, W. F., M. S. Wright, A. H. Lindell, D. M. Harkins, C. Baker-Austin, J. Ravel and R. Stepanauskas (2008). "Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5." Journal of Bacteriology **190**(20): 6779-6794.

Friend, S. H. and T. C. Norman (2013). "Metcalfe's law and the biology information commons." Nature Biotechnology **31**(4): 297-303.

Gibson, G. R. and M. B. Roberfroid (1995). "Dietary Modulation of the Human Colonic Microbiota - Introducing the Concept of Prebiotics." Journal of Nutrition **125**(6): 1401-1412.

Gordon, D. M. and F. FitzGibbon (1999). "The distribution of enteric bacteria from Australian mammals: host and geographical effects." Microbiology-Sgm **145**: 2663-2671.

Guarner, F. and J. R. Malagelada (2003). "Gut flora in health and disease." Lancet **361**(9356): 512-519.

Hammami, R., A. Zouhir, J. Ben Hamida and I. Fliss (2007). "BACTIBASE: a new web-accessible database for bacteriocin characterization." Bmc Microbiology **7**.

Harrison, E. and M. A. Brockhurst (2012). "Plasmid-mediated horizontal gene transfer is a coevolutionary process." Trends in Microbiology **20**(6): 262-267.

Hofer, U. and R. F. Speck (2009). "Disturbance of the gut-associated lymphoid tissue is associated with disease progression in chronic HIV infection." Seminars in Immunopathology **31**(2): 257-277.

Honda, K. and K. Takeda (2009). "Regulatory mechanisms of immune responses to intestinal bacteria." Mucosal Immunology **2**(3): 187-196.

Hongoh, Y. (2010). "Diversity and Genomes of Uncultured Microbial Symbionts in the Termite Gut." Bioscience Biotechnology and Biochemistry **74**(6): 1145-1151.

Jordan, I. K., K. S. Makarova, J. L. Spouge, Y. I. Wolf and E. V. Koonin (2001). "Lineage-specific gene expansions in bacterial and archaeal genomes." Genome Research **11**(4): 555-565.

Kaper, J. B., J. P. Nataro and H. L. T. Mobley (2004). "Pathogenic *Escherichia coli*." Nature Reviews Microbiology **2**(2): 123-140.

Kau, A. L., P. P. Ahern, N. W. Griffin, A. L. Goodman and J. I. Gordon (2011). "Human nutrition, the gut microbiome and the immune system." Nature **474**(7351): 327-336.

Keeling, P. J. and J. D. Palmer (2008). "Horizontal gene transfer in eukaryotic evolution." Nature Reviews Genetics **9**(8): 605-618.

Kurokawa, K., T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi and M. Hattori (2007). "Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes." DNA Research **14**(4): 169-181.

Laing, C., A. Villegas, E. N. Taboada, A. Kropinski, J. E. Thomas and V. P. J. Gannon (2011). "Identification of *Salmonella enterica* species- and subgroup-specific genomic regions using Panseq 2.0." Infection Genetics and Evolution **11**(8): 2151-2161.

Lapierrel, P. and J. P. Gogarten (2009). "Estimating the size of the bacterial pan-genome." Trends in Genetics **25**(3): 107-110.

Leclerc, H., D. A. A. Mossel, S. C. Edberg and C. B. Struijk (2001). "Advances in the bacteriology of the Coliform Group: Their suitability as markers of microbial water safety." Annual Review of Microbiology **55**: 201-234.

Ley, R. E., C. A. Lozupone, M. Hamady, R. Knight and J. I. Gordon (2008). "Worlds within worlds: evolution of the vertebrate gut microbiota." Nature Reviews Microbiology **6**(10): 776-788.

Martin, H. M., B. J. Campbell, C. A. Hart, C. Mpofu, M. Nayar, R. Singh, H. Englyst, H. F. Williams and J. M. Rhodes (2004). "Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer." Gastroenterology **127**(1): 80-93.

McFall-Ngai, M., M. G. Hadfield, T. C. G. Bosch, H. V. Carey, T. Domazet-Loso, A. E. Douglas, N. Dubilier, G. Eberl, T. Fukami, S. F. Gilbert, U. Hentschel, N. King, S. Kjelleberg, A. H. Knoll, N. Kremer, S. K. Mazmanian, J. L. Metcalf, K. Nealson, N. E. Pierce, J. F. Rawls, A. Reid, E. G. Ruby, M. Rumpho, J. G. Sanders, D. Tautz and J. J. Wernegreen (2013). "Animals in a bacterial world, a new imperative for the life sciences." Proceedings of the National Academy of Sciences of the United States of America **110**(9): 3229-3236.

Medini, D., C. Donati, H. Tettelin, V. Masignani and R. Rappuoli (2005). "The microbial pan-genome." Current Opinion in Genetics & Development **15**(6): 589-594.

Muegge, B. D., J. Kuczynski, D. Knights, J. C. Clemente, A. Gonzalez, L. Fontana, B. Henrissat, R. Knight and J. I. Gordon (2011). "Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans." Science **332**(6032): 970-974.

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Ohta, K., D. S. Beall, J. P. Mejia, K. T. Shanmugam and L. O. Ingram (1991). "Genetic-Improvement of *Escherichia coli* for Ethanol Production - Chromosomal Integration of *Zymomonas mobilis* genes encoding Pyruvate Decarboxilase and Alcohol Dehydrogenase II." Applied and Environmental Microbiology **57**(4): 893-900.

Ohta, K., D. S. Beall, J. P. Mejia, K. T. Shanmugam and L. O. Ingram (1991). "Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II." Applied and Environmental Microbiology **57**(4): 893-900.

Oshima, K., H. Toh, Y. Ogura, H. Sasamoto, H. Morita, S. H. Park, T. Ooka, S. Iyoda, T. D. Taylor, T. Hayashi, K. Itoh and M. Hattori (2008). "Complete Genome Sequence and Comparative Analysis of the Wild-type Commensal *Escherichia coli* Strain SE11 Isolated from a Healthy Adult." DNA Research **15**(6): 375-386.

Patyar, S., R. Joshi, D. S. P. Byrav, A. Prakash, B. Medhi and B. K. Das (2010). "Bacteria in cancer therapy: a novel experimental strategy." Journal of Biomedical Science **17**.

Revel, A. T., A. M. Talaat and M. V. Norgard (2002). "DNA microarray analysis of differential gene expression in Borrelial burgdorferi, the Lyme disease spirochete." Proceedings of the National Academy of Sciences of the United States of America **99**(3): 1562-1567.

Sartor, R. B. (2008). "Microbial influences in inflammatory bowel diseases." Gastroenterology **134**(2): 577-594.

Sekirov, I., S. L. Russell, L. C. M. Antunes and B. B. Finlay (2010). "Gut Microbiota in Health and Disease." Physiological Reviews **90**(3): 859-904.

Smillie, C. S., M. B. Smith, J. Friedman, O. X. Cordero, L. A. David and E. J. Alm (2011). "Ecology drives a global network of gene exchange connecting the human microbiome." Nature **480**(7376): 241-244.

Smoot, L. M., J. C. Smoot, M. R. Graham, G. A. Somerville, D. E. Sturdevant, C. A. L. Migliaccio, G. L. Sylva and J. M. Musser (2001). "Global differential gene expression in response to growth temperature alteration in group A Streptococcus." Proceedings of the National Academy of Sciences of the United States of America **98**(18): 10416-10421.

Toh, H., K. Oshima, A. Toyoda, Y. Ogura, T. Ooka, H. Sasamoto, S. H. Park, S. Iyoda, K. Kurokawa, H. Morita, K. Itoh, T. D. Taylor, T. Hayashi and M. Hattori (2010). "Complete Genome Sequence of the Wild-Type Commensal *Escherichia coli* Strain SE15, Belonging to Phylogenetic Group B2." Journal of Bacteriology **192**(4): 1165-1166.